Aprendizaje Estadístico 2025

Lista 2

14.febrero.2025

1. Para un estudio se mide la temperatura en diferentes posiciones del cuerpo de una muestra de personas. Un investigador expresa todas las temperaturas en grados Celcius. Otro investigador convierte primero todas estas temperaturas a grados Fahrenheit. ¿Cómo se relacionan las matrices de covarianza de sus datos?

   Si ambos deciden proyectar en la dirección de máxima varianza, ¿obtendrían las mismas direcciones de proyección? Explique su respuesta.

2. Sea $X = (X_1, X_2, \ldots, X_d)$ una variable aleatoria multidimensional con matriz de covarianza $\mathrm{Cov}(X)$ y esperanza $\mathbb{E}(X) = \mathbf{0}$. Si $l = (l_1, l_2, \ldots, l_d)^T \in \mathbb{R}^d$ es un autovector de $\mathrm{Cov}(X)$ con autovalor λ y $Y = \langle l, X \rangle$, mostrar que $\mathrm{Cov}(Y, X_i) = \lambda\, l_i$.

3. Sea $\mathbb{X} \in \mathbb{R}^{n \times d}$ una matriz de datos arbitraria, y sea $\mathbb{J} = I - \frac{1}{n^2}\mathbf{1}\mathbf{1}^T$ la matriz de $n \times n$ para centrar $\mathbb{X}$. Verificar que

   a) $\mathbb{J}$ es una matriz de proyección.

   b) el vector $\mathbf{1} = (1, \ldots, 1)^T \in \mathbb{R}^n$ es un autovector de $\mathbb{J}$ con autovalor 0.

4. Considerar el conjunto de datos `weather.csv`. Se trata de los promedios mensuales de la temperatura (en Celsius) en 35 estaciones canadienses de monitoreo. El interés es comparar las estaciones entre sí con base en sus curvas de temperatura.

   Considerando las 12 mediciones por estación como un vector $\mathbf{x}$, aplicar un análisis de componentes principales. Como $\mathbf{x}$ representa (un muestreo de) una curva, este tipo de datos se llama datos funcionales.

   - Interpretar y dibujar (como curva) los primeros dos componentes p_1 y p_2. Esto es, graficar $\{(i, p_{1i})\}$ y $\{(i, p_{2i})\}$.
   - Agrupar e interpretar las estaciones en el biplot (tener en mente un mapa de Canadá puede ayudar).

5. A partir de una base de datos con actos delictivos en EE.UU (1970), se construyó la tabla con las correlaciones entre la ocurrencia de 7 clases de delitos, como aparece en la tabla `crimes.dat`. Consideramos cada clase de delito como una observación. Podemos medir la distancia entre dos observaciones como 1 menos su correlación

$$d(X_i, X_j) = 1 - \rho(X_i, X_j), \quad \forall\, i, j$$

   (las correlaciones en la tabla son siempre positivas). Así, la distancia mínima 0 corresponde a correlación máxima 1 entre las variables correspondientes. Encontrar una visualización usando escalamiento multidimensional para estas observaciones y busca una interpretación del eje principal.

6. Entrar al sitio web http://playground.tensorflow.org. En ese sitio aparecen cuatro conjuntos de datos de ejemplo: (1) datos circulares concéntricos, (2) cuatro cuadrantes, (3) dos nubes de puntos gaussianas, y (4) datos en forma de espiral. Construir datos sintéticos en Python que adopten la forma de cada uno de estos conjuntos.

   Para cada uno de los cuatro conjuntos anteriores, entontrar un kernel adecuado para presentar una proyección kernelPCA que modifique la visualización de las nubes de puntos y separe los datos.

7. Construir un sistema de recomendación de filtro colaborativo para el conjunto de películas **movies.csv**, usando como base las recomendaciones de los usuarios en **ratings.csv**, y el método de factoración no-negativa de matrices NNMF. Para ello, puede usar las ideas que vimos en el notebook **recommender.ipynb**.

   Seleccione una métrica apropiada para calcular los vecinos más cercanos entre usuarios, y decida los parámetros de `topNeigh` y `topMovies` a su elección. Muestre ejemplos de varias recomendaciones, y contraste las recomendaciones de su modelo con el top de películas mejor calificadas para un usuario. Repita para 3 usuarios diferentes.
